An Enhanced K Means Clustering using Improved Hopfield Artificial Neural Network and Genetic Algorithm

M.Sakthi, Antony Selvadoss Thanamani

Abstract- Due to the increase in the quantity of data across the world, it turns out to be very complex task for analyzing those data. Categorize those data into remarkable collection is one of the common forms of understanding and learning. This leads to the requirement for better data mining technique. These facilities are provided by a standard data mining technique called Clustering. The key intention of this technique is to categorize a dataset into a set of clusters that contains similar data items, as computed by some distance function. One of the widely used clustering techniques is K-Means clustering. K-Means clustering is very simple and effective for clustering. But, the main disadvantage of this technique is when the large dataset is used for clustering. To overcome this difficulty, various researchers focus on suggesting better alteration in K-Means clustering. This paper provides a new technique to modify K-Means clustering which can result in better performance. For initialization, this paper uses an improved version of Hopfield Artificial Neural Network (HANN) algorithm. Also, the Genetic Algorithm (GA) is in combined with K-Means algorithm. The experimental result indicates that the proposed K-Means clustering algorithm results in better clustering result.

Keywords- K-Means, Genetic Algorithm, Hopfield Artificial Neural Network

I. INTRODUCTION

Due to the advances in sensing and storage tools and extreme growth in the applications like internet search, digital imaging, and video surveillance have produced several high-volume, high-dimensional data sets. Since most of the data are stored digitally in electronic media, they provide high prospective for the growth of automatic data analysis, classification, and retrieval techniques.

Clustering is one of the most accepted techniques utilized for the purpose of data analysis and classification [13]. Cluster analysis is extensively utilized in fields which involve analysis of multivariate data. A search by means of Google Scholar identifies 1,660 entries with the keyword data clustering that comes into view in 2007 alone. This large quantity of data affords the importance of clustering in data analysis. It is especially difficult to list various scientific fields and applications that have utilized clustering technique as well as the thousands of existing techniques.

Manuscript Received on July, 2013.

M.Sakthi, Research Scholar and Head, Department of Computer Science, NGM College, Pollachi, Tamilnadu.

Antony Selvadoss Thanamani, Associate Professor and Head, Department of Computer Science, NGM College, Pollachi, Tamilnadu.

The key intention of data clustering is to detect the natural classification of a set of patterns, points, or objects. Webster defines cluster analysis as "a statistical classification technique for identifying whether the individuals of a population fall into various categories by means quantitative comparisons of multiple characteristics". The another definition of clustering is: Provided a representation of n objects, determine K groups according to the measure of similarity like similarities among objects in the same group are high whereas the similarities between objects in different groups are low.

The main advantages of using the clustering algorithms are:

- Compactness of representation.
- Fast, incremental processing of new data points.
- Clear and fast identification of outliers.

The widely used clustering technique is K-Means clustering [7, 9]. This is because K-Means is very simple to implement and also it is effective in clustering. But K-Means clustering will lack performance when large dataset is involved for clustering. Various researchers try to overcome those difficulties by their own ideas [14, 16]. To deal with this problem, this paper suggests a better technique. For initializing the clustering, this paper uses Hopfield Artificial Neural Network (HANN) technique. In this paper, an improved version of the HANN is utilized. In the new algorithm, the problem is considered as a minimization of an energy function constructed of a cost-term as a sum of squared errors. To ensure the convergence of the network, the minimization is achieved with a step function permitting the network to reach its stability in a pre-specified period of time. The K-Means algorithm is also modified by using the Genetic Algorithm.

II. RELATED WORK

Zhe *et al.,* [1] presents an improved K-Means clustering algorithm. K-means algorithm [10] is broadly applied in spatial clustering. The mean value of all the cluster centroid in this method is considered as the Heuristic information; therefore it has few demerits like sensitive to the initial centroid and instability. The improved clustering technique referred to the best clustering centroid that are searched while the optimization process is performed in clustering centroid. This enhances the searching accuracy around the best centroid and improves the strength of the clustering technique.

Hai-xiang Guo *et al.,* [2] put forth an Improved Genetic k-means Algorithm for Optimal Clustering. The value of k must be known in advance in the traditional k-means approach [10, 11]. It is very tough to

confirm the value of k accurately in advance. The author proposed an enhanced genetic k-means clustering (IGKM) and builds a fitness function defined as a product of three factors, maximization of which guarantees the formation of a small number of compact clusters with large separation between at least two clusters. Finally, the experiments are conducted on two artificial and three real-life data sets that compare IGKM with other traditional methods like k-means algorithm, GA-based technique and genetic k-means algorithm (GKM) by inter-cluster distance (ITD), inner-cluster distance (IND) and rate of separation exactness. From the experimental observation, it is clear that IGKM reach the optimal value of k with high accuracy.

Yanfeng Zhang *et al.*, [3] proposed an Agglomerative Fuzzy K-means clustering method with automatic selection of cluster number (NSS-AKmeans) approach for learning optimal number of clusters and for providing significant clustering results. High density areas can be detected by the NSS-AKmeans and from these centers the initial cluster centers with a neighbor sharing selection approach can also be determined. Agglomeration Energy (AE) factor is proposed in order to choose a initial cluster for representing global density relationship of objects. Moreover, in order to calculate local neighbor sharing relationship of objects, Neighbors Sharing Factor (NSF) is used. Agglomerative Fuzzy k-means clustering algorithm is then utilized to further merge these initial centers to get the preferred number of clusters and create better clustering results. Experimental observations on several data sets have proved that the proposed clustering approach was very significant in automatically identifying the true cluster number and also providing correct clustering results.

Xiaoyun Chen *et al.*, [4] described a GK-means: an efficient K-means clustering algorithm based on grid. Clustering analysis is extensively used in several applications such as pattern recognition, data mining, statistics etc. K-means approach, based on reducing a formal objective function, is most broadly used in research. But, user specification is needed for the k number of clusters and it is difficult to choose the effective initial centers. It is also very susceptible to noise data points. In this paper, the author mainly focuses on option the better initial centers to enhance the quality of k-means and to minimize the computational complexity of k-means approach. The proposed GK-means integrates grid structure and spatial index with k-means clustering approach [15]. Theoretical analysis and experimental observation show that the proposed approach performs significantly with higher efficiency.

Trujillo *et al.*, [5] proposed a combining K-means and semivariogram-based grid clustering approach. Clustering is widely used in various applications which include data mining, information retrieval, image segmentation, and data classification. A clustering technique for grouping data sets that are indexed in the space is proposed in this paper. This approach mainly depends on the k-means clustering technique and grid clustering. K-means clustering [8, 12] is the simplest and most widely used approach. The main disadvantage of this approach is that it is sensitive to the selection of the initial partition. Grid clustering is extensively used for grouping data that are indexed in the space. The main aim of the proposed clustering approach is to eliminate the high sensitivity of the k-means clustering approach to the starting conditions by using the available spatial information. A semivariogram-based grid clustering technique is used in this approach. It utilizes the spatial correlation for obtaining the bin size. The author combines this approach with a conventional k-means clustering technique as the bins are constrained to regular blocks while the spatial distribution of objects is irregular. An effective initialization of the k-means is provided by semivariogram. From the experimental results, it is clearly observed that the final partition protects the spatial distribution of the objects.

Huang *et al.*, [6] put forth the automated variable weighting in k-means type clustering that can automatically estimate variable weights. A novel approach is introduced to the k-means algorithm to iteratively update variable weights depending on the present partition of data and a formula for weight calculation is also proposed in this paper. The convergency theorem of the new clustering algorithm is given in this paper. The variable weights created by the approach estimates the significance of variables in clustering and can be deployed in variable selection in various data mining applications where large and complex real data are often used. Experiments are conducted on both synthetic and real data and it is found from the experimental observation that the proposed approach provides higher performance when compared the traditional k-means type algorithms in recovering clusters in data.

III. METHODOLOGY

In the proposed clustering technique, the number of clusters are determined using the Hopfield Artificial Neural Network (HANN), and then the Genetic K-means algorithm is applied to determine the final solution. The process involved in the proposed system is provided in figure 1.

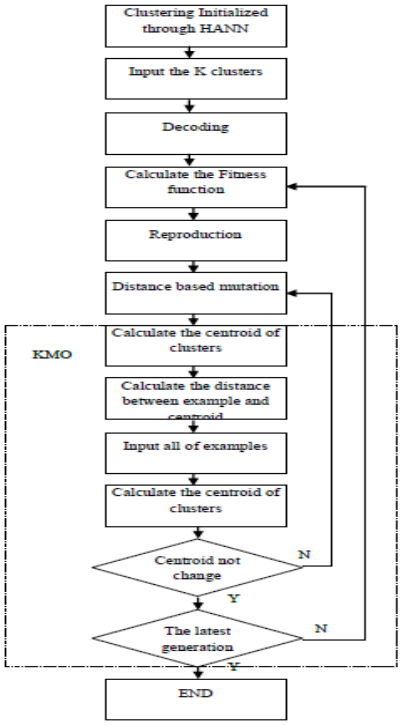

Figure 1: Proposed Clustering Technique

This study proposes an algorithm which proceeds as follows:

- It uses the HANN to find the initial clusters.
- The distance-based mutation is used to escape local solutions and to find the global solution
- It uses the GKA adopted K-Means Operator (KMO) for faster convergence.

Hopfield Artificial Neural Network (HANN)

The HANN structure contains a single layer indicating a grid of N x M neurons with every column indicating a class and every row indicating a pixel. Every neuron acts as both input and output neurons at the same instance. Actually neurons under every class grasp the probability that the equivalent pixel belongs to this class. N indicates the size of the provided data and M indicates the number of classes. The network is build in order to cluster the data without learning phase according to the compactness of every class computed with the help of distance measure (R_{kl}) among the kth pixel and the centroid of class l. The clustering solution is prepared as a partition of N points among M classes in a manner that the assignment of the data reduces the cost-term of the energy (error) function:

$$E = \frac{1}{2}\sum_{k=1}^{N}\sum_{l=1}^{M} R_{kl}^{n} V_{kl}^{2}$$

Where R_{kl} is indicated as the distance measure among the kth pixel and the centroid of class l, and defined as below:

$$R_{kl} = ||X_k - \bar{X}_l||$$

Where X_k represents the feature value (intensity value) of the kth point and X_l indicates the centroid value of class l, and defined as below:

$$\bar{X}_l = \frac{\sum_{k=1}^{N} X_k V_{kl}}{n_l}$$

Where n_l represents the number of data points in class l. In this case n=2 which indicates that the energy is defined as sum-squared error. V_{kl} indicates the output of the k_lth neuron. The rule adopted here is winner-takes-all learning rule, where the input-output function for the kth row (to allocate a label m to the kth pixel) is provided by:

$$V_{kl}(t+1) = 1 \qquad if\ U_{kl} = Max[U_{kl}, 1]$$
$$V_{kl}(t+1) = 0 \qquad otherwise$$

The minimization resulted by utilizing hopfield neural network and by solving a set of motion equations satisfying:

$$\frac{\partial U_i}{\partial t} = -\mu(t)\frac{\partial E}{\partial V_i}$$

Where U_i and V_i represents the input and output of the ith neuron respectively, μ (t) indicates a scalar positive function of time that finds the length of the step to be consider in the direction of the vector d = −∇E(V) . The suitable choice of the step μ (t) is some thing of an art, experimentation and a familiarity with a provided class of optimization problems are usually necessary to determine the best function. μ (t) function utilized in this paper:

$$\mu(t) = t^{*}(T_s - t)$$

Where t represents the iteration step and T_s indicates the pre-specified convergence time. HANN segmentation algorithm can be summarized in the following steps:

1. Initialize the input of neurons to random values.
2. Apply the input-output function (V_{kl}) defined above, to obtain the new output values for each neuron, establishing the assignment of data points to classes. The class membership probabilities grow or diminish in a winner-takes-all style as a result of contention between classes. In winner-takes-all model, the neuron with the highest input value fires and takes the value 1, and all remaining neurons take the value 0.
3. Compute the centroid (X l) as defined above, for each class l.
4. Compute the energy function (E) as defined above,
5. Update the inputs (U_i) using the following equation, learning occurs here; when neuron input weights are adjusted in an attempt to reduce the output error.

$$U_{kl}(t+1) = U_{kl}(t) + \frac{\partial U_i}{\partial t}$$

6. Repeat from step 2 until t = Ts. This process iteratively modifies the pixel label assignments to reach a near optimal final segmentation map.

After the number of clusters is determined by using HANN, Genetic algorithm based K-Means (GKA) clustering algorithm is utilized to clusters the data.

The GKA coding phase is reserved in this research. The offspring is created according to the best fitness function, distance-based mutation and K-means operator in the population. The GKA steps are as below.

The key intention of the clustering technique considered here is to partition a collection of n provided pattern; every pattern is a vector of dimension d, into K groups, such that this partition reduces the Total Within Cluster Variance (TWCV), which is defined as below.

Let $\{x_i, i=1, 2,\ldots, n\}$ be the set of n patterns; let x_{ij} represents the jth feature of x_i; define i=1, 2,…, n and k=1, 2,…, K.

$$W_{ik} = \begin{cases} 1, & if\ ith\ pattern\ belong\ to\ kth\ cluster \\ 0, & otherwise \end{cases}$$

Then, the matrix W= $[w_{ij}]$ has the properties that

$$W_{ik} \in \{0,1\}., \quad \sum_{k}^{K} W_{ij} = 1$$

The centroid of cluster k is

$$c_{kj} = \frac{\sum_{i=1}^{n} W_{ik} x_{ij}}{\sum_{i=1}^{n} W_{ik}}$$

The within-cluster variation of kth cluster is defined as

$$S^{(k)}(W) = \sum_{i=1}^{n} W_{ik} \sum_{i=1}^{d} (x_{ij} - c_{kj})^2$$

and the TWCV is defined as

$$S(W) = \sum_{k=1}^{K} S^{(k)}(W) = \sum_{k=1}^{K} \sum_{i=1}^{n} W_{ik} \sum_{j=1}^{d} (x_{ij} - c_{kj})^2$$

The objective is to minimize S(W), i.e.,

$$S(W)^* = \min_{W} \{S(W)\}$$

Coding System

An usual manner of coding such W into a string, s_w, is to consider a chromosome of length n and to permit all the allele in the chromosome to acquire value from {1, 2,…, K}. In this situation, all the allele belonging to a pattern and its value indicates the cluster number to which the respective pattern fit in.

Initialize population

The starting population is chosen randomly. Every allele in the population can be initialized to a cluster number randomly chosen from the uniform distribution in the values {1, 2, . . ., K}. This is rejected by assigning p, the maximum integer which is less than n/K, randomly selected data points to every cluster and the rest of points to randomly selected clusters.

Fitness function selection

The selection operator chose a chromosome by means of random manner from the previous population based on the distribution provided by

$$P(S_i) = \frac{F(S_i)}{\sum_{i=1}^{N} F(S_i)}$$

where $F(s_i)$ indicates the fitness value of the string s_i in the population. This type of random choosing concern to the roulette wheel; conversely, the string s_w's fitness function value depends on TWCV. The lesser the S(W), higher the $f(s_w)$. In addition, $\bar{f}$ and r symbolize the present population $f(s_w)$'s average and standard error, correspondingly. Additionally, c is a value in [1,3]. $f(s_w)$ and $F(s_w)$ are given as:

$$f(S_w) = -S(W), \quad gf(S_w) = f(S_w) - (\bar{f} - c\sigma)$$

$$f(S_w) = -S(W), \qquad if\ g(S_w) \geq 0$$

Mutation

Mutation modifies an allele value based on the distances of the cluster centroids from the respective data point. To utilize the mutation operator to the allele $s_w(i)$ related to pattern X_i, let $d_j = d(X_i, C_j)$ be the Euclidean distance among X_i and C_j (described as average of jth dimension). Subsequently, the allele is replaced with a value selected randomly from the below distribution:

$$p_j = P_r\{S_w(i) = j\} = \frac{c_m d_{max} - d_j}{\sum_{i=1}^{K} (c_m d_{max} - d_i)}$$

$$where\ c_m \geq 1\ and\ d_{max} = max_j\{d_j\}$$

K-Means Operator

Initialization:

Assume a random option such as i (i=1, 2. . ., m) and cluster j (j=1, 2. . ., c), set j = 1 and k=1. (M_j^0: illustration allocated to the jth cluster, I_j^0: index of the illustration assigned the jth cluster)

Algorithms:

Step 1: Compute the centroid of cluster j.

$$Z_j^k = \begin{cases} \frac{1}{M_j^{k-1}} \sum_{i \in I_j^{k-1}} x_i, & if\ M_j^{k-1} > 0 \end{cases}$$

Step 2: Compute the distance among every illustation and centroid.

$$j_1^k = \sum_{j=1}^{c} \sum_{i \in I_j^{k-1}} \| x_i - Z_j^k \|^2$$

Step 3: Compute the new assignment, illustration i assigned to the j* cluster (i.e., w_{ij}*= 1, w_{ij}=0, j=1,…, c; j ≠ j*)

$$\| x_i - z_{j^*}^k \|^2 \leq \| x_i - Z_j^k \|^2, \quad j = 1, \ldots\ldots, c, j \neq j^*$$

If the equal sign is true, then terminate. If i <m, then perform the step 3 again and let i = i + 1; otherwise all illustrations are assigned the jth cluster and all indices of illustrations are assigned to the jth cluster, go to step 4.

Step 4: Compute the distance among illustrations and centroid.

$$J_2^k = \sum_{j=1}^{c} \sum_{i \in I_j^{k-1}} \| x_i - Z_j^k \|^2$$

Step 5: If $\left|J_1^k - j_2^k\right| < \varepsilon$,, then stop, else j = j + 1 and k = E_{k+1} go to step 1.

The proposed clustering will result in better optimization and cluster separation.

IV. EXPERIMENTAL RESULTS

The proposed technique is experimented using the two benchmark datasets which are Iris and Wine Dataset from the UCI machine learning Repository [17]. The experiments are all performed on a GENX computer with 2.6 GHz Core (TM) 2 Duo processors using MATLAB version 7.5.

Experiment with Iris Dataset

The Iris flower data set (Fisher's Iris data set) is a multivariate data set. The dataset comprises of 50 samples from each of three species of Iris flowers (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from every sample; they are the length and the width of sepal and petal, in centimeters. Based on the combination of the four features, Fisher has developed a linear discriminant model to distinguish the species from each other. It is used as a typical test for many classification techniques. The proposed method is tested first using this Iris dataset. This database has four continuous features consisting of 150 instances: 50 for each class.

To evaluate the efficiency of the proposed approach, this technique is compared with the existing K-Means algorithm. The Mean Square Error (MSE) of the centers $MSE = \sqrt{||v_c - v_t||^2}$ where v_c is the computed center and v_t is the true center. The cluster centers found by the proposed K-Means are closer to the true centers, than the centers found by K-Means algorithm. The mean square error for the four cluster centers for the two approaches are presented in table I. The resulted execution time for the proposed and standard K-Means algorithms is provided in figure 2 and the accuracy resulted is provided in figure 3.

TABLE I
MEAN SQUARE ERROR VALUE OBTAINED FOR THE THREE CLUSTERS IN THE IRIS DATASET

	SOM+K-Means	HANN +Genetic K- Means
Cluster 1	0.2765	0.1825
Cluster 2	0.3025	0.1998
Cluster 3	0.2514	0.1724

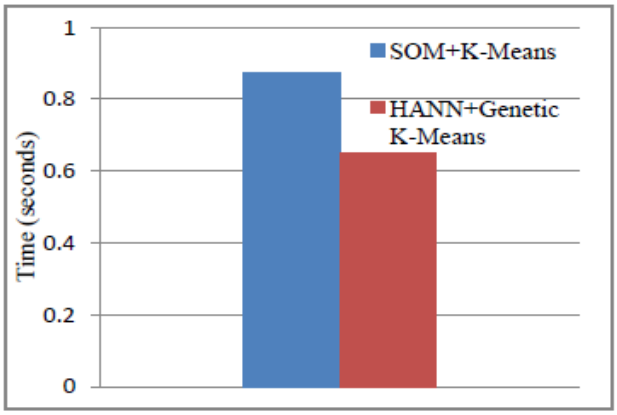

Figure 2: Execution Time for Iris Dataset

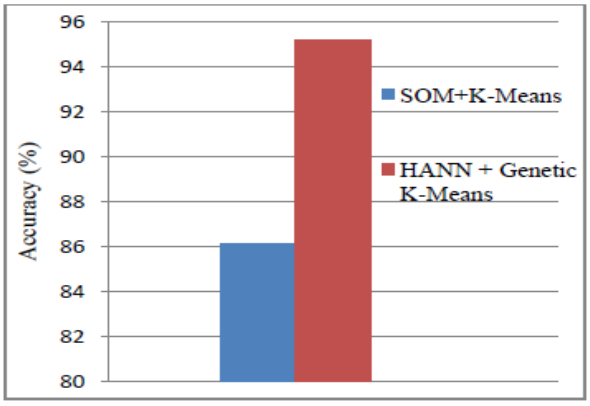

Figure 3: Accuracy for Iris Dataset

Experiment with Wine Dataset

The wine dataset is the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis established the quantities of 13 constituents found in each of the three types of wines. The classes 1, 2 and 3 have 59, 71 and 48 instances respectively. There are totally 13 Number of Attributes.

The MSE value for the three clusters is presented in Table II. The resulted execution for the proposed and standard K-Means algorithms is provided in figure 4 and the resulted accuracy is presented in figure 5.

TABLE II
MEAN SQUARE ERROR VALUE OBTAINED FOR THE THREE CLUSTERS IN THE WINE DATASET

	SOM+K-Means	HANN +Genetic K- Means
Cluster 1	0.3581	0.2915
Cluster 2	0.4125	0.3105
Cluster 3	0.1824	0.1624

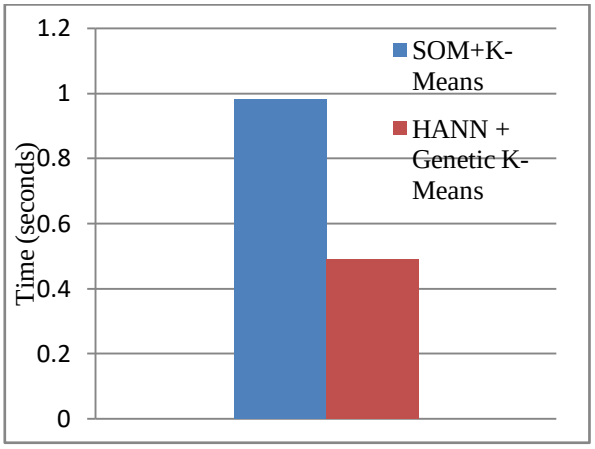

Figure 4: Execution Time for Wine Dataset

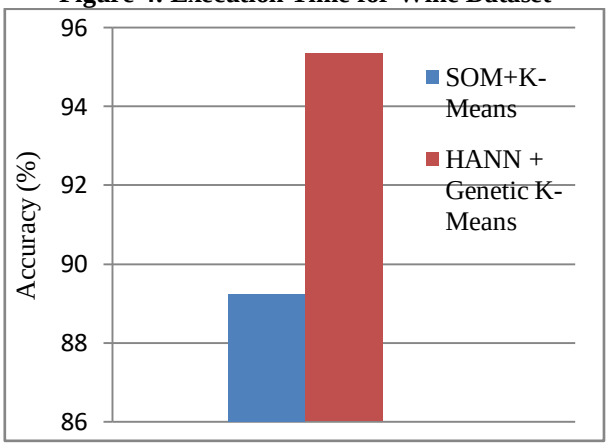

Figure 5: Accuracy for Wine Dataset

From the experimental observations it can be found that the proposed approach produces better clusters than the existing approach. The MSE value is highly reduced for both the dataset. This represents the better accuracy for the proposed approach. Also, the execution time is reduced when compared to the existing approach. This is true in both the dataset.

V. CONCLUSION

The increase in the number of data world wide leads to the requirement for the better analyzing technique for better understanding of data. One of the most essential modes of understanding and learning is categorizing data into reasonable groups. This can be achieved by a famous data mining technique called Clustering. Clustering is nothing but separating the given data into particular groups according to the separation among the data points. This will helps in better understanding and analyzing of the vast data. One of the widely used clustering is K-Means clustering because it is simple and efficient. But it lacks accuracy of classification when large data are used in clustering. So the K-Means clustering needs to be improved to suit for all kinds of data. This paper enhanced the K-Means clustering by using Hopfield Artificial Neural Network and Genetic Algorithm. The initial parameters for K-Means clustering is provided by using HANN and the K-Means algorithm itself is modified by using GA. The experimental results suggest that the proposed technique results in better classification when compared to conventional techniques.

REFERENCES

1. Zhang Zhe, Zhang Junxi and Xue Huifeng, "Improved K-Means Clustering Algorithm", Congress on Image and Signal Processing, Vol. 5, Pp. 169-172, 2008.
2. Hai-xiang Guo, Ke-jun Zhu, Si-wei Gao and Ting Liu, "An Improved Genetic k-means Algorithm for Optimal Clustering", Sixth IEEE International Conference on Data Mining Workshops, Pp. 793-797, 2006.
3. Yanfeng Zhang, Xiaofei Xu and Yunming Ye, "NSS-AKmeans: An Agglomerative Fuzzy K-means clustering method with automatic selection of cluster number", 2nd International Conference on Advanced Computer Control, Vol. 2, Pp. 32-38, 2010.
4. Xiaoyun Chen, Youli Su, Yi Chen and Guohua Liu, "GK-means: an Efficient K-means Clustering Algorithm Based on Grid", International Symposium on Computer Network and Multimedia Technology, Pp. 1-4, 2009.
5. Trujillo, M., Izquierdo, E., "Combining K-means and semivariogram-based grid clustering", 47th International Symposium, Pp. 9-12, 2005.
6. Huang, J.Z., Ng, M.K., Hongqiang Rong and Zichen Li, "Automated variable weighting in k-means type clustering", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 5, Pp. 657-668, 2005.
7. Yi Hong and Sam Kwong "Learning Assignment Order of Instances for the constrained k-means clustering algorithm" IEEE Transactions on Systems, Man, and Cybernetics, Vol 39, No 2. April, 2009.
8. I. Davidson,M. Ester and S.S. Ravi, "Agglomerative hierarchical clustering with constraints: Theoretical and empirical results", in Proc. of Principles of Knowledge Discovery from Databases, PKDD 2005.
9. Wagstaff, Kiri L., Basu, Sugato, Davidson, Ian "When is constrained clustering beneficial, and why?" National Conference on Aritficial Intelligence, Boston, Massachusetts 2006.
10. Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schrodl "Constrained K-means Clustering with Background Knowledge" ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning, 2001.
11. I. Davidson, M. Ester and S.S. Ravi, "Efficient incremental constrained clustering". In Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2007, August 12-15, San Jose, California, USA.
12. I. Davidson, M. Ester and S.S. Ravi, "Clustering with constraints: Feasibility issues and the K-means algorithm", in proc. SIAM SDM 2005, Newport Beach, USA.
13. D. Klein, S.D. Kamvar and C.D. Manning, "From Instance-Level constraintes to space-level constraints: Making the most of Prior Knowledge in Data Clustering", in proc. 19th Intl. on Machine Learning (ICML 2002), Sydney, Australia, July 2002, p. 307-314.
14. N. Nguyen and R. Caruana, "Improving classification with pairwise constraints: A margin-based approach", in proc. of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD'08).
15. K. Wagstaff, C. Cardie, S. Rogers and S. Schroedl, "Constrained Kmeans clustering with background knowledge", in: Proc. Of 18th Int. Conf. on Machine Learning ICML'01, p. 577 - 584.
16. Y. Hu, J. Wang, N. Yu and X.-S. Hua, "Maximum Margin Clustering with Pairwise Constraints", in proc. of the Eighth IEEE International Conference on Data Mining (ICDM) , 253-262, 2008.
17. Merz C and Murphy P, UCI Repository of Machine Learning Databases, Available: ftp://ftp.ics.uci.edu/pub/machine-Learning-databases.
18. Text Documents Clustering using Genetic Algorithm and Discrete Differential Evolution. *International Journal of Computer Applications* 43(1):16-19, April 2012. Published by Foundation of Computer Science, New York, USA. BibTeX